How High School Teachers Develop Tests and How AI Could Help

Yuan "Charles" Cui[1][0000−0002−2681−6441], Ph.D. Candidate in Computer Science
Mia Cheng[1], B.S. Candidate in Computer Science
Matthew Kay[1][0000−0001−9446−0419], Associate Professor of Computer Science and Communication Studies
Fumeng Yang[2][0000−0002−8401−2580], Assistant Professor of Computer Science

[1] Northwestern University, Evanston IL 60208, USA
[2] University of Maryland, College Park MD 20742, USA

Abstract. Tests play a critical role in the educational ecosystem, but developing tests can be difficult. To understand teachers' current practices and identify opportunities for AI to support, we interview 13 high school teachers across different subjects and schools in the USA. Using thematic analysis, we capture teachers' diverse and complex test development processes in a conceptual model, which consists of three categories of inputs and four stages of actions: teachers consider the inputs from external rules and standards, their interactions with students, and available materials; they move between the stages of refining test requirements, searching for relevant materials, assessing and restructuring tests, as well as creating (also adapting or copying) an item. From this model, we identify four categories of challenges teachers face: (1) developing high-quality items; (2) overcoming technical barriers in item development; (3) navigating test-level constraints; (4) working with imperfect AI. We discuss opportunities to address these challenges with AI, grounding our discussion in existing work from AI in education, creativity support tools, sensemaking, and AI literacy. Our work reveals a broad space of opportunities for deeper exploration and innovation in AI for test development.

Keywords: Human-Centered Design · AI for Test Development · Qualitative Study

1 Introduction

Tests play a crucial role in K–12 education. For teachers, tests allow them to assess the effectiveness of instructional methods and refine teaching strategies [9, 33]. For students, tests help them identify knowledge gaps and reinforce learning [52, 75]. Well-crafted tests ensure quality and accountability in education by providing consistent, meaningful feedback to all stakeholders.

AI models, particularly large language models (LLMs), have shown promise in generating test items (questions) [13, 22, 23, 31, 58], the building blocks of tests. LLMs can be used to create items across different subjects (e.g., social science [57], mathematics [50]) that measure lower-level skills but struggle with higher-level abilities [17, 50]. Human oversight and interaction might improve

the item generation process, but most existing work relies on fully automated, end-to-end approaches with little human involvement.

Some researchers have taken a human-centered approach to item development using AI. For example, Wang et al. studied how college instructors design quiz questions about assigned readings and identified how NLP can help [71], which led to the development of a tool that supports this specific task [37]. While some lessons from these works may transcend school levels, K–12 teachers may have different constraints, question types, subject matter, and test development practices. If we can better understand K–12 teachers' existing test development processes, we can learn how to best integrate AI into their practices.

To that end, we conduct a semi-structured interview study with 13 high school teachers in the USA. We focus on high school because it is a period in K–12 education where students encounter frequent, high-stakes testing [5]. Following a thematic analysis [12] of the interviews, we:

- **Develop a conceptual model** to capture teachers' iterative test development processes. Our model includes three categories of inputs and four stages of action. Teachers consider these inputs and iteratively move between the stages to simultaneously develop the test and refine test requirements.
- **Identify four categories of challenges** teachers face in test development: (1) developing high-quality items (e.g., combining multiple concepts into one item); (2) overcoming technical barriers of item development (e.g., making a data visualization); (3) navigating test-level constraints (e.g., following state standards on science education); (4) working with imperfect AI (e.g., evaluating and refining potentially erroneous AI-generated test items).
- **Discuss how to address these challenges** by forging **new** connections between them and a broad body of literature—including work in AIED, creativity support tools, sensemaking, and AI literacy. While prior work has proposed targeted solutions for some of these challenges, we focus on uncovering new connections that may lead to novel approaches for AI support in test development.

Our conceptual model, insights, and discussion outline a broad space for integrating AI into test development, with teachers at the center. Future work can build on our results, taking a human-centered approach to developing solutions that better support K–12 teachers. Our work presents opportunities for deeper exploration and innovation in AI for test development.

2 Related Work

Test Development. Researchers in education, psychology, and cognitive science have proposed measurement theories and models to guide test development, such as constructing detailed test blueprints [16, 69], estimating item parameters (e.g., difficulty) [6, 25, 28], and assessing validity and reliability with expert panels and statistical analysis [14, 61]. Some argue that models of cognition and learning can inform educational test development, making tests more effective in measuring student understanding [49]. However, translating such research into practice is not

easy [43, 49]. In reality, teachers rarely adopt scientific principles when creating tests, as these theories often lack relevance to their day-to-day instruction [43, 66]. Consequently, existing work in test development theory provides insufficient practical guidance for how teachers currently develop tests.

AI for Item Generation and Evaluation. Rapid advancements in AI, particularly LLMs, have spurred significant interest in AI application for item generation and evaluation [13, 22, 58, 62]. Researchers have attempted to generate different types of items (e.g., multiple-choice items [23, 29, 31]), create items across different subjects (e.g., social science [57], mathematics [50]), and automatically evaluate item quality [38, 44, 45]. While current AI models have shown promise in generating items across diverse contexts [58], they often struggle to produce items that can measure higher cognitive abilities [17, 50]. To enhance the capabilities of LLMs in item generation, researchers have used fine-tuning [36] and retrieval-augmented generation [29], but these methods are often limited by the quality of external knowledge bases and datasets. Most existing work also lacks mechanisms for integrating human oversight and interaction into the item generation process. Without such mechanisms, there are substantial concerns about content quality, alignment with pedagogical objectives, and teachers' adoption in practice [18]. Therefore, to ensure quality and adoption, it is important to consider a human-centered approach to design AI-based solutions for test development.

Human-Centered Design in AIED. Existing guidelines for human-AI systems in education highlight teacher participation to ensure instructional relevance and better serve students [18, 24, 34, 35]. In line with these guidelines, Wang et al. conducted a study with college instructors on how AI can help make reading questions based on academic papers [71]. This led to the development of ReadingQuizMaker, which helps college instructors with this specific task [37]. While these works demonstrate the potential of human-centered approaches in AI for item development, much remains to be done—particularly in deepening our understanding of K–12 teachers' test development processes—to better design AI tools.

3 Methodology

Existing literature reveals an insufficient understanding of high school teachers' current practices and subsequently how AI can be effectively integrated into them. Taking a human-centered approach, we ask the following research questions (RQ):

RQ1: How do high school teachers develop tests?

RQ2: (a) What are the challenges high school teachers face in test development, and (b) how could AI help teachers address these challenges?

To answer these questions, we conducted a semi-structured interview study, which allows us to obtain nuanced perspectives and deep insights from participants [21].

Recruitment We first recruited from schools with which the research team and their institutions had an existing partnership program. We emailed department chairs and teachers of the mathematics, science, and social studies departments to to describe our study and invite them to participate. As we recruited participants,

we also used snowball sampling, asking participants if they knew other teachers who fit characteristics that our current sample lacked (e.g., teachers with fewer years of experience). Before interviewing a participant, we emailed them the consent form and scheduled brief online meetings to answer any questions about our study. All participants signed the consent form after this preliminary meeting.

Participants We obtained 13 participants (6 female and 7 male) from 8 schools (6 public and 2 private) in 3 states. All participants are 18 years or older, speak fluent English, and were actively teaching at the time of the interview. We had 5 teachers from mathematics departments, 6 from science departments, and 2 from social studies departments. Their teaching experience ranged from 2 to 28 years. The study was approved by our Institutional Review Board, and each participant received a $40 electronic gift card as compensation.

Interview Protocol The first author conducted one-on-one semi-structured Zoom interviews with participants. Before the interviews started, we obtained consent to record both the audio and video of the meeting and gave them the option of not turning on their video. During the interview, we used seed questions[3] to facilitate the conversation. Each interview took about an hour.

Thematic Analysis The first author transcribed and anonymized all audio recordings. Two authors (including the first author) conducted a thematic analysis [12] on all 13 transcripts. The two authors open-coded all the transcripts. The first author then created an affinity diagram [39] based on the initial coding to extract themes. A third author then reviewed and commented on themes, and these three authors iteratively refined, split, and merged the themes through several rounds of discussion to reach a consensus [42].[4]

4 Current Processes: Inputs & Stages

Teaching is a demanding profession. Teachers are often pressed for time and face numerous responsibilities. Our participants typically teach 3–5 classes per year, each with 5–30 students. In addition to lesson planning, instruction, grading, feedback, and parent communication, they must regularly develop tests—a complex, time-consuming task requiring both expertise and labor. To understand teachers' processes, we develop a conceptual model in Fig. 1 that highlights **three categories of inputs**$^{\downarrow}$ and **four stages**$^{\circlearrowleft}$ **of actions** in test development.

4.1 Inputs$^{\downarrow}$ to Test Development

We categorize the inputs$^{\downarrow}$ to test development into three main areas: (1) external rules and standards$^{\downarrow}$, (2) interactions with current students$^{\downarrow}$, and (3) available materials$^{\downarrow}$. These inputs shape teachers' test requirements$^{\checkmark}$ and guide their decisions throughout the test development process.

[3] Two authors created a list of questions based on our research questions, which was then revised by two other authors to ensure clarity and avoid leading questions.

[4] As the initial coding is not the product but the process to generate themes, we do not compute measures such as inter-rater reliability [42].

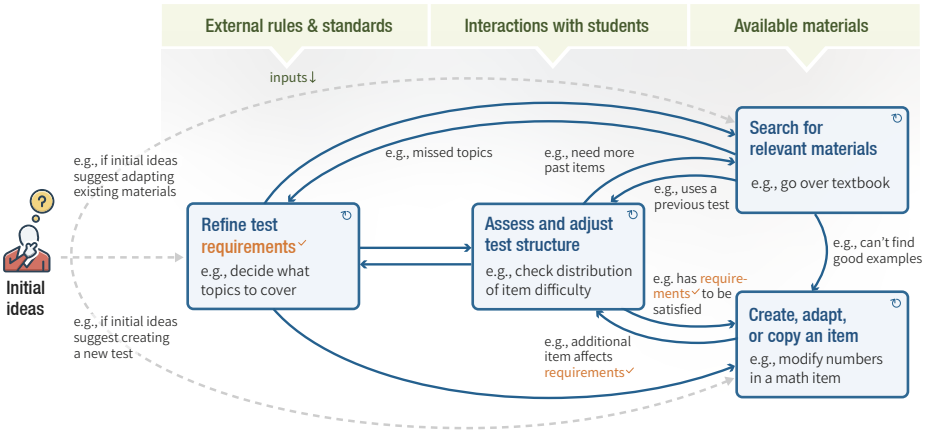

Fig. 1. Our conceptual model depicting teachers' processes of test development. In this model, a teacher has some initial ideas about the test requirements (✓). They then start (⇢) and move between the stages (↻) to simultaneously develop the test and refine test requirements, while considering various inputs (↓).

External Rules & Standards↓ Teachers do not have complete autonomy in designing tests; they must adhere to rules and standards from curricula, schools, districts, or states. In Illinois, for instance, the Next Generation Science Standards (NGSS) define what students should know and how they should demonstrate understanding in K–12 science education [73]. For example, P8 created biology test items emphasizing data analysis to align with NGSS's focus on cross-cutting skills. Standardized, high-stakes exams also exert a strong influence: Advanced Placement (AP) is a program that offers university-level curricula and exams to high school students in the USA and Canada [10]. Teachers often mimic the AP exam formats in their tests <P1,3,4,9>. As P4 noted, "*we do try to mirror the AP exam ... that exam will have some multiple choice and some free response.*" Beyond standardized content, logistical constraints such as limited exam time and quick grading requirements further restrict test development: "*I've gotta turn grades around within 24 hours, so I can't do a bunch of open-ended questions ... I have to write a multiple-choice test*" <P5>.

Teacher's Interactions with Current Students↓ Teaching is a continuous interaction in which teachers observe students' progress through classroom activities and formative assessments (e.g., quizzes), then adapt tests to suit students' specific strengths, struggles, backgrounds, and expectations. As P11 noted, "*I'm spending a lot of my job watching how my kids do the problem ... when I make the tests, I know I wanna ask the type of questions that they've been tripped up on ... I wanna see that they're getting past that.*" These multifaceted interactions help teachers shape their tests to reflect learners' needs and precisely gauge understanding.

Available Materials↓ Creating tests is time-intensive, and teachers rarely start from scratch. They typically draw from various resources, including self-curated item banks, colleague-shared materials, official curricula (e.g., released past AP

exams), and online sources (e.g., teacher forums). While these materials often prove useful, their quantity, quality, and relevance vary, influencing how much teachers need to adapt or create new content. As P5 noted, "*I'm not sure where they're sourcing their questions ... their diagrams are really outdated ... I'll have like an idea of what question I want to ask, and then I'll go to Google Images ... then I will just re-tailor the question to fit that diagram.*" Nonetheless, teachers frequently rely on these materials as a key input in test development.

4.2 Stages$^{\circlearrowleft}$ of Test Development

Teachers rarely define all test requirements$^{\checkmark}$—topics to cover, difficulty levels, item formats, test length, and alignment with students' experiences$^{\downarrow}$, etc <P1,9,12,13>—at the outset. Instead, they engage in a **dual, iterative** process: they (1) develop the test while they simultaneously (2) refine its requirements. Some begin with a vague sense of test topics, refining those topics as they explore available materials$^{\downarrow}$ <P1,2,12>. Others draw on semi-developed requirements$^{\checkmark}$ from past courses and focus on creating or adapting items, making occasional adjustments to test requirements$^{\checkmark}$ for current students$^{\downarrow}$ <P8,9>. Below, we split this dual process into four stages: one for refining test requirements, and three for developing the test.

Refine Test Requirements$^{\circlearrowleft}$ Teachers begin the test development process with a set of initial ideas about their test requirements$^{\checkmark}$, which evolve as they develop the test in the dual process outlined above. Updating the requirements$^{\checkmark}$ may prompt teachers to search$^{\circlearrowleft}$ from available resources$^{\downarrow}$ or create$^{\circlearrowleft}$ new items, and to assess and restructure$^{\circlearrowleft}$ the test for any gaps. For instance, after reviewing past exam performance, P8 "*either change*[*s*] *the questions or* [*gives*] *students a little bit more support and scaffolding*" to improve a new test. Teachers continue refining requirements throughout the process until the final test meets their goals.

Search for Relevant Materials$^{\circlearrowleft}$ If teachers' initial ideas suggest that they should develop the test by adapting available materials$^{\downarrow}$, they may begin by searching for relevant content, such as past tests, homework, quizzes, textbook problem sets, curriculum-prescribed item banks, or online resources: "*I can look through* [*old textbooks*] *to quickly glance and see ... that looks like it matches what I'm trying to assess from my students*" <P1>. From here, teachers may transition to create, adapt, or copy$^{\circlearrowleft}$—adapting or copying if they find relevant materials or creating a new item if they do not. In this stage, teachers may also spot gaps that prompt them to assess and restructure the test$^{\circlearrowleft}$ or refine test requirements$^{\circlearrowleft}$: "*I kind of look through* [*my notes and grade book*] *and say, am I forgetting anything? Is there anything that we taught that we spent time on that I'd like to test these students on?*" <P12>. Teachers often return to the search stage repeatedly throughout the test development process to seek relevant materials.

Create, Adapt, or Copy an Item$^{\circlearrowleft}$ Teachers may also begin by creating new items if their initial ideas suggest the need for a brand new test. This is common for first-time teachers who may not have sufficient available materials to adapt

<P2,12> or those who prefer novelty and worry about test leakage[5] <P4,13>. For instance, P13 tried to create new items that differ from students' expectations$^{\downarrow}$: "*I try to make ... a question that presents itself a little bit differently than exactly what everybody thinks is coming.*" Teachers also adapt or copy past items. This is often less laborious, especially for experienced teachers. If they find a helpful item by searching$^{\circlearrowleft}$ from their available materials$^{\downarrow}$, they can copy it to the work-in-progress test or adapt it to fit the new test; for example, P7 said: "*I've never given the same test ever ... at the very least I take a previous question and change some variables to it so they can do it in a ... different context.*" As teachers add items to the test, they regularly assess and restructure$^{\circlearrowleft}$ it to decide what to do next.

Assess & Restructure Test$^{\circlearrowleft}$ If teachers have a past test from searching$^{\circlearrowleft}$ or a work-in-progress test, they may assess and restructure it based on their current requirements$^{\checkmark}$, such as coverage of topics, distribution of difficulty, adherence to external standards$^{\downarrow}$, or characteristics of current students$^{\downarrow}$: "*I tweak* [*old tests*] *every year ... I make sure that I cover the standards and the things that I've stressed within that semester*" <P9>. If teachers find the current test does not satisfy their requirements$^{\checkmark}$, they may search$^{\circlearrowleft}$ for materials or create, adapt, or copy$^{\circlearrowleft}$ items to address the gaps: "*I pulled up the last couple years* [*of final exams*] *... we got further this year ... So the final exam had to account for an assessment on cell rests. So we also looked at some cell rest questions that we were gonna add to the final to make it a little beefier there*" <P7>. Teachers revisit this stage until all requirements$^{\checkmark}$ are met, which concludes the test development process.

5 Challenges and Opportunities

As shown in Sec. 4, test development is a dual, iterative process in which teachers consider multiple inputs while simultaneously developing tests and refining test requirements$^{\checkmark}$. Beyond crafting individual items, teachers must also balance topics and difficulty levels. Drawing on our conceptual model, we highlight four categories of challenges in the stages of teachers' processes. We then identify potential AI-based solutions by forging new connections between these challenges and a broad body of literature—including work in AIED, creativity support tools, sensemaking, and AI literacy. While previous research has proposed targeted solutions for some challenges we identified, we focus on uncovering new connections that may lead to novel, alternative approaches for AI to support test development.

5.1 Developing High-Quality Items

Developing high-quality items$^{\circlearrowleft}$ is a difficult task. Teachers must navigate many factors—from integrating multiple concepts into one item to balancing clarity and sophistication—each posing its own challenges.

[5] Test leakage is the release or sharing of test materials that could compromise the fairness and integrity of the test; e.g., if past years' students share tests.

C1: Combining & Scaffolding Teachers often design test items that combine multiple concepts to challenge students' problem-solving skills <P1,3-7>. Crafting such items requires juggling different materials$^{\downarrow}$—often without a database to efficiently query past items that meet specific requirements$^{\checkmark}$ <P3>—and creatively generating divergent ideas to link different concepts: "*I'm trying to cover ... different concepts in one problem ... I'm pulling problems from multiple sources ... that's a creative process because I'm trying to put this bigger picture together*" <P1>. In contrast, teachers scaffold complex items by breaking them into parts, guiding students through problem-solving while accommodating different ability levels <P2,4,5,7,8>. For instance, P7 scaffolds an item by starting with "what" questions and gradually progresses to "how" and "why." This approach demands significant effort and creativity, as teachers must deeply understand students' progress and struggles$^{\downarrow}$ to "*step into that student hat*" <P8> to craft effective scaffolding.

C2: Contextual Adaptation Test items should reflect students' experiences$^{\downarrow}$ to promote equity and relevance <P8,10-12>, which requires teachers to identify students' interests as well as social and cultural backgrounds, and then carefully integrate them into item development. For instance, P8 tries to create items that are grounded in the shared experiences among all her students, ensuring that the test is equitable. Teachers also consider their students' backgrounds to make items practical and relevant beyond the classroom. For example, P10 designed a test item for math students about to enter the age of car ownership: "*I made all my students buy a car* [*in a math problem*] *... just to see as soon as you drive it off the lot, it's not worth what you just paid for it. Not only were they learning about exponential decay and growth, but they were also learning how to budget ... credit scores and things like that.*" These adaptations enhance learning, but even minor contextual tweaks—such as incorporating students' names or local geography into an item—demand significant time and effort <P11>.

C3: Validity A *valid* test item accurately measures its corresponding skills [11]. Ensuring validity takes many forms. Teachers strive to design items that target specific skills without requiring extraneous knowledge—for example, in a math problem not about fractions, a teacher might use integers to minimize student errors due to incorrect operations on fractions <P1,10-12>. Effective distractors are key to properly assessing students and identifying exactly how they make mistakes in multiple-choice items, but can be difficult to create <P3,4,12>. Additionally, teachers work to balance clarity and sophistication in item wording; as P13 noted, it can be hard to be clear without "*spoonfeeding*" students. Achieving validity requires meticulous attention to detail to ensure items assess the intended abilities.

We identified the following opportunities for the challenges above:

O1: Tagging Items for Easy Search When developing high-quality items, teachers often need to search$^{\circlearrowleft}$ for past items$^{\downarrow}$ that meet specific requirements$^{\checkmark}$ (e.g., topics, skills), whether they are looking for items on particular concepts to combine or scaffold (C1), or adapting the context (C2) of an item to connect with students' needs. However, they often don't have items tagged for easy filtering and querying. LLMs can automatically generate and tag knowledge components (specific skills

or competencies) for items, and Moore et al. developed a clustering algorithm to group items that assess similar knowledge components [46]. This approach could be integrated into a bank of teacher's materials to automatically label each item.

O2: Creativity Support for Item Development Developing an item is a creative process, in which a teacher first generates divergent ideas, such as how concepts could be combined (C1), how an item could be scaffolded (C1), or what possible distractors or question phrasings might improve the validity (C3) of a multiple-choice item. They then analyze and synthesize these possibilities, converging onto a single design. Generating many ideas and considering them all at once is challenging `<P4-7,9>`. Idea generation techniques and interaction paradigms from creativity support tools [26] can help users generate, view, and synthesize ideas in the divergent-convergent process of various creative tasks [65, 79]. For example, Suh et al. developed a framework for structured idea generation with LLMs: given a task prompt (e.g., write a story about dogs), the framework prompts the LLM to identify key dimensions (e.g., plot complexity) and their possible values (e.g., "suspenseful"), then generates multiple responses based on combinations of these dimensions [64]. Reza et al. created an interface that supports rapid, visually structured exploration of variations in writing tasks, helping users consider multiple variations simultaneously as they refine their writing [53]. Such interaction techniques could help teachers generate and synthesize the divergent ideas necessary to overcome difficult challenges in item design, such as combining topics, scaffolding items, or improving item phrasing (validity).

O3: Customization in Item Development Teachers often need to integrate students' backgrounds and learning progress$^{\downarrow}$ into item development (e.g., scaffolding (C1) and contextual adaptation (C2)). Researchers have demonstrated the potential to create customized educational content using AI [1, 4, 27]. For example, Abolnejadian et al. developed context-setting prompt templates to generate content tailored to introductory programming students' educational backgrounds [2]. A system built around this method could allow teachers to enter information about their students (e.g., interests, skills they have mastered, what they struggle with) and then use prompt templates to produce customized items (C2). In addition, recent work has developed generative student agents that perform similarly to real students with similar profiles [38, 48]. These methods could help teachers "*step into that student hat*" `<P8>` when developing scaffolded items, using AI student agents to assess whether an item is structured in a way that provides sufficient support for students of a variety of ability levels (C1).

5.2 Overcoming Technical Barriers in Item Development

C4: Creating Special Symbols & Visual Representations Special symbols are prevalent in STEM[6], where precise notation is key. Visual representations—such as free body diagrams in physics `<P5>` or choropleth maps in geography `<P2>`—are prevalent across disciplines to convey complex scenarios and data. They are

[6] Science, technology, engineering and mathematics.

frequently used in items that test students' domain knowledge or data analysis skills. Creating them often requires specialized software and tools. However, teachers often lack the requisite technical expertise or do not have the budget and time to purchase and learn how to use these tools, necessitating workarounds. For example, P10 struggled to add mathematical symbols in Google Docs, so he resorted to manually adjusting tests after printing or verbally informing students what he had intended to write. Similarly, P11 created bar and pie charts in Excel (a tool she knew), screenshotted them, and inserted them into Google Docs. She repeated this process each time she needed to adjust the question or chart.

O4: Lowering Technical Barriers for Symbols and Visuals AI solutions can lower the technical barrier for creating special symbols and visual representations (C4). For example, we could allow users to input symbols using other modalities, e.g., by describing what they need in natural language (using LLMs to translate to LaTeX syntax) or through sketching interfaces [41, 47, 56, 63, 67]. AI-based visualization authoring and generation tools could also help teachers without programming experience create visual representations for test items by specifying their requirements in natural language [17, 20, 70].

5.3 Navigating Test-Level Constraints

C5: Following Rules and Standards External rules and standards$^{\downarrow}$ are often convoluted <P3-5>; it is hard for teachers to translate them into actionable requirements$^{\checkmark}$ for test development: "*the way NGSS is presented is bizarre, because it's presented in really formal, frankly overly academic language that I think is really difficult for many teachers to decipher*" <P5>. Similarly, P3 and P4 noted that AP's Course and Exam Descriptions span hundreds of pages, making it hard to extract practical guidelines. Frequent updates to these standards impose additional burdens, such as the need to continually revise tests and the uncertainty in aligning tests with current standards.

O5: Summarizing Rules and Standards Language models (e.g., GPT [3] and BERT [19]) have improved text summarization techniques, converting lengthy and jargon-filled documents into digestible summaries [68, 78]. They also power educational Q&A tools, such as virtual tutors [59, 80]. These models could be adapted to generate concise overviews of external standards and enable interactive Q&A conversation about how these standards affect test development (C5). They could be used to generate descriptions of changes in standards and identify the discrepancy between teachers' current tests and updated standards. In turn, teachers could more easily create or adapt items$^{\circlearrowleft}$, refine test requirements$^{\circlearrowleft}$ or assess and restructure$^{\circlearrowleft}$ tests based on changes in standards over the years.

5.4 Working with Imperfect AI

Participants' experiences with AI varied—from never using it to using it regularly to generate presentation outlines. Some have even used AI to create tests (e.g., practice exams <P4> and unit tests <P12>). However, all teachers who used AI encountered difficulties and expressed mixed satisfaction with their experience.

C6: Understanding and Using AI Understanding the capabilities of AI could help teachers set realistic expectations and integrate AI into their work, building trust in the process <P2,13>. Yet, effective use of AI requires skills [60, 77] that many teachers lack, such as crafting effective prompts: "*You have to keep figuring out how to ask* [*ChatGPT*] *to get what you want ... I'm not perfect at that*" <P6>. This need for specialized training, which we also noted in special symbols & visual representations (C4), adds to teachers' already heavy workload, and access to high-quality training resources is often limited.

C7: Evaluating and Refining AI Output All of the teachers who used AI tools (e.g., ChatGPT) felt it was important to evaluate and refine AI-generated content. In their evaluation of an AI-generated test, these teachers tried to answer every item themselves to ensure the questions were clearly stated and the answers were correct <P3-5,9-12>. They also refer to existing materials to check whether a generated item is similar to a previous item <P3>. One serious issue they found is that generated items involving calculations often exhibit hallucinations <P9,10,12>. Dealing with hallucinations is challenging because some calculation errors are hard to spot quickly <P10>. At the test level, teachers also had to assess↻ if the test requirements✓ were satisfied (e.g., balanced coverage of topics and difficulty), and sometimes needed to reorder (restructure)↻ the items on a test for a desirable structure (e.g., progression of difficulty). For example, P5 would consult his printed test outline, review each ChatGPT-generated item, and note on the outline which topics each item covers, ensuring that all areas are evenly represented. However, conversational AI tools such as ChatGPT do not sufficiently support teachers' evaluation and refinement. Teachers often had to scroll back and forth in the chat interface to find an item, and switched between multiple tools repeatedly (e.g., Google Docs, ChatGPT) to edit, label, or remove items <P12>. It was also inconvenient to give AI feedback on specific items in the chat interface, resulting in a labor-intensive process <P9,10,12>.

We identified the following opportunities for the challenges above:

O6: AI Literacy Training Researchers have noted the demand for AI literacy training for teachers [15, 76]. Casal-Otero et al. argued that to help teachers understand and use AI it is critical to identify their current perception and knowledge of AI, provide updated professional development, and involve them in the design of new curricula where AI plays a role [15]. Researchers have created training programs outside of the teaching context; for example, Ma et al. developed a training paradigm that improves novices' ability to articulate requirements in prompting by having them develop digital games through prompting [40]. Such paradigms can be adapted to incorporate various tasks in test development as targeted training for teachers (C6).

O7: Sensemaking for Test Evaluation and Refinement Evaluating a test↻ is both an *information exploration* [74] and a *sensemaking* [51] task: to understand test quality, teachers scan different items, assess their individual quality and relationships with each other, and decide what to keep, adapt↻, or remove. Interface designs and interaction techniques from this literature—such as semantic

zooming [7, 72], filtering [30], navigation [8, 54], and note-taking [55]—could help teachers evaluate a test (C7). For example, LLMs could tag generated items by topics and format, which teachers could use as filters to evaluate items on a test; as teachers examine individual items, AI-driven recommender systems [32] could suggest similar items from teachers' existing materials and allow teachers to compare them side by side; to get an overview of topic coverage, teachers can use semantic zoom to quickly see the high-level information of all items on the test.

6 Discussion and Conclusion

An overarching theme that emerged from our study is the limited time teachers have for developing tests, coupled with the critical need for trust in the tools they use. If AI-based solutions require extra steps or slow down an already tight workflow, teachers may not feel the added value justifies the effort. Trust in these tools can be built over time through the ability to evaluate and refine AI-generated content; however, if this iterative process is too time-consuming, teachers simply cannot invest the necessary effort. As a result, prolonged evaluation can quickly erode trust, leading teachers to reject AI tools that seem to complicate their established, reliable practices. For AI-based test development tools to gain traction, developers must design them to integrate seamlessly into established workflows, reducing the burden of test development without adding complexity, and building trust by demonstrating immediate, tangible benefits. Otherwise, teachers—especially those with limited experience or time—are unlikely to adopt solutions that fail to offer both reliable performance and clear benefits.

Through our interviews, we also found teachers' test development practices vary based on subject area, experience level, and school type. For example, STEM teachers often spend significant time ensuring numeric accuracy and correctness in test items, making the development process more labor-intensive. In contrast, social studies teachers frequently employ open-ended items that lack a single right answer, which can mean less time spent on crafting the test but substantially greater effort required for grading. New teachers generally devote considerable time to creating tests from scratch and finding reliable resources, whereas experienced teachers typically rely on established routines to adapt existing materials. In addition, resource availability and administrative expectations can vary widely between public and private schools, shaping the ways teachers approach test development. Given these differences, it is essential that AI solutions for test development are tailored to the target audience. Designers and developers should identify their audience early on, account for the particularities of that group, and design solutions that directly address their unique challenges and requirements.

Despite the valuable insights gained from interviewing teachers of diverse backgrounds, this study has several limitations. First, the small sample size of 13 teachers—all from the USA—may not fully represent the wide spectrum of cultural, institutional, and curricular factors that influence test development. Additionally, our recruitment method may have introduced self-selection bias, as teachers who chose to participate could have been more inclined toward

AI adoption than the broader educator population. We also faced particular challenges in recruiting new teachers, who often experience the most significant burden in test development and have far less time available. Future research should engage larger and more diverse teacher populations, employ iterative, human-centered design processes to develop AI-based tools, and conduct rigorous field evaluations to validate that these tools meaningfully support test development. We hope these human-centered approaches will open up new opportunities for integrating AI into test development, empowering teachers to enhance assessments, and ultimately improving student learning outcomes.

References

1. Abbes, F., Bennani, S., Maalel, A.: Generative AI and Gamification for Personalized Learning: Literature Review and Future Challenges. SN Comput. Sci. **5**(8) (Dec 2024). https://doi.org/10.1007/s42979-024-03491-z
2. Abolnejadian, M., Alipour, S., Taeb, K.: Leveraging ChatGPT for Adaptive Learning through Personalized Prompt-based Instruction: A CS1 Education Case Study. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. CHI EA '24, Association for Computing Machinery (2024). https://doi.org/10.1145/3613905.3637148
3. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 Technical Report (2024), https://arxiv.org/abs/2303.08774
4. Adair, A., Pedro, M.S., Gobert, J., Segan, E.: Real-Time AI-Driven Assessment and Scaffolding that Improves Students' Mathematical Modeling during Science Investigations. In: Wang, N., Rebolledo-Mendez, G., Matsuda, N., Santos, O.C., Dimitrova, V. (eds.) Artificial Intelligence in Education. pp. 202–216. Springer Nature Switzerland, Cham (2023)
5. Amrein, A.L., Berliner, D.C.: High-Stakes Testing, Uncertainty, and Student Learning. Education Policy Analysis Archives **10**, 18 (Mar 2002). https://doi.org/10.14507/epaa.v10n18.2002
6. Association, A.E.R., Association, A.P., on Measurement in Education, N.C.: Standards for Educational and Psychological Testing. American Educational Research Association (2014)
7. Bederson, B.B., Hollan, J.D.: Pad++: A Zooming Graphical Interface for Exploring Alternate Interface Physics. In: Proceedings of the 7th Annual ACM Symposium on User Interface Software and Technology. p. 17–26. UIST '94, Association for Computing Machinery (1994). https://doi.org/10.1145/192426.192435
8. Benyon, D.: The New HCI? Navigation of Information Space. Know.-Based Syst. **14**(8), 425–430 (Dec 2001). https://doi.org/10.1016/S0950-7051(01)00135-6
9. Black, P.J.: Formative and Summative Assessment by Teachers. Studies in Science Education **21**(1), 49–97 (1993). https://doi.org/10.1080/03057269308560014, https://doi.org/10.1080/03057269308560014
10. Board, C.: Advance Placement (AP) Program (2025), https://ap.collegeboard.org/
11. Borsboom, D., Mellenbergh, G.J., Van Heerden, J.: The Concept of Validity. Psychological review **111**(4), 1061 (2004). https://doi.org/10.1037/0033-295X.111.4.1061
12. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qualitative Research in Psychology **3**(2), 77–101 (2006). https://doi.org/10.1191/1478088706qp063oa

13. Bulathwela, S., Muse, H., Yilmaz, E.: "Scalable Educational Question Generation with Pre-trained Language Models". In: Wang, N., Rebolledo-Mendez, G., Matsuda, N., Santos, O.C., Dimitrova, V. (eds.) Artificial Intelligence in Education. pp. 327–339. Springer Nature Switzerland, Cham (2023)
14. Carmines, E.G.: Reliability and Validity Assessment. Sage University Paper Series on Quantitative Applications in the Social Sciences (1979). https://doi.org/10.4135/9781412985642
15. Casal-Otero, L., Catala, A., Fernández-Morante, C., Taboada, M., Cebreiro, B., Barro, S.: AI literacy in K-12: a systematic literature review. International Journal of STEM Education **10**(1), 29 (2023). https://doi.org/10.1186/s40594-023-00418-7
16. Cohen, R.J., Schneider, W.J., Tobin, R.: Psychological Testing and Assessment. McGraw Hill LLC, New York, NY, 10th edn. (2022)
17. Cui, Y., Ge, L.W., Ding, Y., Harrison, L., Yang, F., Kay, M.: Promises and Pitfalls: Using Large Language Models to Generate Visualization Items. IEEE Transactions on Visualization and Computer Graphics **31**(1), 1094–1104 (2025). https://doi.org/10.1109/TVCG.2024.3456309
18. Cukurova, M., Miao, X., Brooker, R.: Adoption of Artificial Intelligence in Schools: Unveiling Factors Influencing Teachers' Engagement. In: Wang, N., Rebolledo-Mendez, G., Matsuda, N., Santos, O.C., Dimitrova, V. (eds.) Artificial Intelligence in Education. pp. 151–163. Springer Nature Switzerland, Cham (2023)
19. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: North American Chapter of the Association for Computational Linguistics (2019), https://api.semanticscholar.org/CorpusID:52967399
20. Dibia, V.: LIDA: A Tool for Automatic Generation of Grammar-Agnostic Visualizations and Infographics using Large Language Models. In: Bollegala, D., Huang, R., Ritter, A. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations). pp. 113–126. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-demo.11
21. DiCicco-Bloom, B., Crabtree, B.F.: The Qualitative Research Interview. Medical Education **40**(4), 314–321 (2006). https://doi.org/10.1111/j.1365-2929.2006.02418.x
22. Do, H., Lee, G.G.: "Aspect-Based Semantic Textual Similarity for Educational Test Items". In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education. pp. 344–352. Springer Nature Switzerland, Cham (2024)
23. Dutulescu, A., Ruseti, S., Iorga, D., Dascalu, M., McNamara, D.S.: "Beyond the Obvious Multi-choice Options: Introducing a Toolkit for Distractor Generation Enhanced with NLI Filtering". In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education. pp. 242–250. Springer Nature Switzerland, Cham (2024)
24. Felix, C.V.: The Role of the Teacher and AI in Education, vol. 33, pp. 33–48. Emerald Publishing Limited (2025/02/20 2020). https://doi.org/10.1108/S2055-364120200000033003, https://doi.org/10.1108/S2055-364120200000033003
25. Ferketich, S.: Focus on psychometrics. Aspects of item analysis. Research in Nursing & Health **14**(2), 165–168 (1991). https://doi.org/10.1002/nur.4770140211
26. Frich, J., MacDonald Vermeulen, L., Remy, C., Biskjaer, M.M., Dalsgaard, P.: Mapping the Landscape of Creativity Support Tools in HCI. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. p. 1–18. CHI '19, Association for Computing Machinery (2019). https://doi.org/10.1145/3290605.3300619

27. Ghosh, A., Tschiatschek, S., Devlin, S., Singla, A.: Adaptive Scaffolding in Block-Based Programming via Synthesizing New Tasks as Pop Quizzes. In: Rodrigo, M.M., Matsuda, N., Cristea, A.I., Dimitrova, V. (eds.) Artificial Intelligence in Education. pp. 28–40. Springer International Publishing, Cham (2022)
28. Haladyna, T.M., Rodriguez, M.C.: Developing and Validating Test Items. Routledge (2013). https://doi.org/10.4324/9780203850381
29. Hang, C.N., Wei Tan, C., Yu, P.D.: MCQGen: A Large Language Model-Driven MCQ Generator for Personalized Learning. IEEE Access **12**, 102261–102273 (2024). https://doi.org/10.1109/ACCESS.2024.3420709
30. Hearst, M.A.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics (1992), https://aclanthology.org/C92-2082/
31. Hwang, K., Wang, K., Alomair, M., Choa, F.S., Chen, L.K.: "Towards Automated Multiple Choice Question Generation and Evaluation: Aligning with Bloom's Taxonomy". In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education. pp. 389–396. Springer Nature Switzerland, Cham (2024)
32. Khanal, S.S., Prasad, P.W.C., Alsadoon, A., Maag, A.: A Systematic Review: Machine Learning Based Recommendation Systems for E-Learning. Education and Information Technologies **25**(4), 2635–2664 (2020). https://doi.org/10.1007/s10639-019-10063-9
33. Kibble, J.: Best practices in summative assessment. Advances in Physiology Education **41**(1), 16–25 (2017). https://doi.org/10.1152/advan.00116.2016, https://doi.org/10.1152/advan.00116.2016
34. Kim, J.: Leading teachers' perspective on teacher-AI collaboration in education. Education and Information Technologies **29**(7), 8693–8724 (2024). https://doi.org/10.1007/s10639-023-12109-5, https://doi.org/10.1007/s10639-023-12109-5
35. Kim, J., Lee, H., Cho, Y.H.: Learning design to support student-AI collaboration: perspectives of leading teachers for AI in education. Education and Information Technologies **27**(5), 6069–6104 (2022). https://doi.org/10.1007/s10639-021-10831-6, https://doi.org/10.1007/s10639-021-10831-6
36. Lamsiyah, S., El Mahdaouy, A., Nourbakhsh, A., Schommer, C.: Fine-Tuning a Large Language Model with Reinforcement Learning for Educational Question Generation. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education. pp. 424–438. Springer Nature Switzerland, Cham (2024)
37. Lu, X., Fan, S., Houghton, J., Wang, L., Wang, X.: ReadingQuizMaker: A Human-NLP Collaborative System that Supports Instructors to Design High-Quality Reading Quiz Questions. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23, Association for Computing Machinery (2023). https://doi.org/10.1145/3544548.3580957
38. Lu, X., Wang, X.: Generative Students: Using LLM-Simulated Student Profiles to Support Question Item Evaluation. In: Proceedings of the Eleventh ACM Conference on Learning @ Scale. p. 16–27. L@S '24, Association for Computing Machinery (2024). https://doi.org/10.1145/3657604.3662031
39. Lucero, A.: Using Affinity Diagrams to Evaluate Interactive Prototypes. In: Abascal, J., Barbosa, S., Fetter, M., Gross, T., Palanque, P., Winckler, M. (eds.) Human-Computer Interaction – INTERACT 2015. pp. 231–248. Springer International Publishing, Cham (2015)

40. Ma, Q., Peng, W., Yang, C., Shen, H., Koedinger, K., Wu, T.: What Should We Engineer in Prompts? Training Humans in Requirement-Driven LLM Use (2024), https://arxiv.org/abs/2409.08775
41. Mathpix: Mathpix (2025), https://mathpix.com/
42. McDonald, N., Schoenebeck, S., Forte, A.: Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. Proc. ACM Hum.-Comput. Interact. **3**(CSCW) (Nov 2019). https://doi.org/10.1145/3359174
43. McMillan, J.H.: Understanding and Improving Teachers' Classroom Assessment Decision Making: Implications for Theory and Practice. Educational measurement: Issues and practice **22**(4), 34–43 (2003)
44. Moore, S., Costello, E., Nguyen, H.A., Stamper, J.: An Automatic Question Usability Evaluation Toolkit. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education. pp. 31–46. Springer Nature Switzerland, Cham (2024)
45. Moore, S., Nguyen, H.A., Bier, N., Domadia, T., Stamper, J.: Assessing the Quality of Student-Generated Short Answer Questions Using GPT-3. In: Hilliger, I., Muñoz-Merino, P.J., De Laet, T., Ortega-Arranz, A., Farrell, T. (eds.) "Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption". Springer International Publishing, Cham (2022)
46. Moore, S., Schmucker, R., Mitchell, T., Stamper, J.: Automated Generation and Tagging of Knowledge Components from Multiple-Choice Questions. In: Proceedings of the Eleventh ACM Conference on Learning @ Scale. p. 122–133. L@S '24, Association for Computing Machinery (2024). https://doi.org/10.1145/3657604.3662030
47. MyScript: MyScript (2025), https://www.myscript.com/
48. Nair, I.J., Tan, J., Su, X., Gere, A., Wang, X., Wang, L.: Closing the Loop: Learning to Generate Writing Feedback via Language Model Simulated Student Revisions. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) "Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing". pp. 16636–16657. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.emnlp-main.928
49. Pellegrino, J.W., Chudowsky, N., Glaser, R. (eds.): Knowing What Students Know: The Science and Design of Educational Assessment. The National Academies Press, Washington, DC (2001). https://doi.org/10.17226/10019
50. Pham, P.V.L., Duc, A.V., Hoang, N.M., Do, X.L., Luu, A.T.: ChatGPT as a Math Questioner? Evaluating ChatGPT on Generating Pre-university Math Questions. In: Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing. p. 65–73. SAC '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3605098.3636030
51. Pirolli, P., Card, S.: The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis. In: Proceedings of International Conference on Intelligence Analysis. vol. 5, pp. 2–4. McLean, VA, USA (2005)
52. Raupach, T., Brown, J., Anders, S., Hasenfuss, G., Harendza, S.: Summative Assessments Are More Powerful Drivers of Student Learning Than Resource Intensive Teaching Formats. BMC Medicine **11**(1), 61 (2013). https://doi.org/10.1186/1741-7015-11-61, https://doi.org/10.1186/1741-7015-11-61
53. Reza, M., Laundry, N.M., Musabirov, I., Dushniku, P., Yu, Z.Y.M., Mittal, K., Grossman, T., Liut, M., Kuzminykh, A., Williams, J.J.: ABScribe: Rapid Exploration & Organization of Multiple Writing Variations in Human-AI Co-Writing

Tasks using Large Language Models. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. CHI '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3613904.3641899
54. Rivlin, E., Botafogo, R., Shneiderman, B.: Navigating in Hyperspace: Designing a Structure-Based Toolbox. Commun. ACM **37**(2), 87–96 (Feb 1994). https://doi.org/10.1145/175235.175242
55. Roy, N., Torre, M.V., Gadiraju, U., Maxwell, D., Hauff, C.: Note the Highlight: Incorporating Active Reading Tools in a Search as Learning Environment. In: Proceedings of the 2021 Conference on Human Information Interaction and Retrieval. p. 229–238. CHIIR '21, Association for Computing Machinery (2021). https://doi.org/10.1145/3406522.3446025
56. Sakshi, Kukreja, V.: Machine Learning and Non-machine Learning Methods in Mathematical Recognition Systems: Two Decades' Systematic Literature Review. Multimedia Tools and Applications **83**(9), 27831–27900 (2024). https://doi.org/10.1007/s11042-023-16356-z
57. Scaria, N., Chenna, S.D., Subramani, D.: "How Good are Modern LLMs in Generating Relevant and High-Quality Questions at Different Bloom's Skill Levels for Indian High School Social Science Curriculum?". In: Kochmar, E., Bexte, M., Burstein, J., Horbach, A., Laarmann-Quante, R., Tack, A., Yaneva, V., Yuan, Z. (eds.) "Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)". pp. 1–10. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024), https://aclanthology.org/2024.bea-1.1/
58. Scaria, N., Dharani Chenna, S., Subramani, D.: Automated Educational Question Generation at Different Bloom's Skill Levels Using Large Language Models: Strategies and Evaluation. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education. pp. 165–179. Springer Nature Switzerland, Cham (2024)
59. Schmucker, R., Xia, M., Azaria, A., Mitchell, T.: Ruffle&Riley: From Lesson Text to Conversational Tutoring. In: Proceedings of the Eleventh ACM Conference on Learning @ Scale. p. 547–549. L@S '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3657604.3664719
60. Sclar, M., Choi, Y., Tsvetkov, Y., Suhr, A.: Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=RIu5lyNXjT
61. Shillingburg, W.: Understanding validity and reliability in classroom, school-wide, or district-wide assessments to be used in teacher/principal evaluations. Journal of behavioral education **10**(4), 205–212 (2016)
62. Shimmei, M., Bier, N., Matsuda, N.: "Machine-Generated Questions Attract Instructors When Acquainted with Learning Objectives". In: Wang, N., Rebolledo-Mendez, G., Matsuda, N., Santos, O.C., Dimitrova, V. (eds.) Artificial Intelligence in Education. pp. 3–15. Springer Nature Switzerland, Cham (2023)
63. Sketch2scheme: Sketch2scheme (2025), https://sketch2scheme.com/
64. Suh, S., Chen, M., Min, B., Li, T.J.J., Xia, H.: Luminate: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. CHI '24, Association for Computing Machinery (2024). https://doi.org/10.1145/3613904.3642400
65. Sultanum, N., Srinivasan, A.: DATATALES: Investigating the use of Large Language Models for Authoring Data-Driven Articles. In: 2023 IEEE Visualization and Visual

Analytics (VIS). pp. 231–235. IEEE Computer Society, Los Alamitos, CA, USA (oct 2023). https://doi.org/10.1109/VIS54172.2023.00055
66. Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J.M., Milligan, S., Selwyn, N., Gašević, D.: Assessment in the age of artificial intelligence. Computers and Education: Artificial Intelligence **3**, 100075 (2022). https://doi.org/10.1016/j.caeai.2022.100075
67. Truong, T.N., Nguyen, C.T., Zanibbi, R., Mouchère, H., Nakagawa, M.: A survey on handwritten mathematical expression recognition: The rise of encoder-decoder and GNN models. Pattern Recognition **153**, 110531 (2024). https://doi.org/10.1016/j.patcog.2024.110531
68. Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.B., Aali, A., Bluethgen, C., Pareek, A., Polacin, M., Reis, E.P., Seehofnerová, A., et al.: Adapted Large Language Models Can Outperform Medical Experts in Clinical Text Summarization. Nature Medicine **30**(4), 1134–1142 (2024). https://doi.org/10.1038/s41591-024-02855-5
69. Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., Herrera-Seda, C.: Authentic Assessment: Creating A Blueprint for Course Design. Assessment & Evaluation in Higher Education **43**(5), 840–854 (2018). https://doi.org/10.1080/02602938.2017.1412396
70. Wang, C., Thompson, J., Lee, B.: Data Formulator: AI-Powered Concept-Driven Visualization Authoring. IEEE Transactions on Visualization and Computer Graphics **30**(1), 1128–1138 (2024). https://doi.org/10.1109/TVCG.2023.3326585
71. Wang, X., Fan, S., Houghton, J., Wang, L.: Towards Process-Oriented, Modular, and Versatile Question Generation that Meets Educational Needs. In: Carpuat, M., de Marneffe, M.C., Meza Ruiz, I.V. (eds.) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 291–302. Association for Computational Linguistics, Seattle, United States (Jul 2022). https://doi.org/10.18653/v1/2022.naacl-main.22
72. Ware, C., Osborne, S.: Exploration and virtual camera control in virtual three dimensional environments. In: Proceedings of the 1990 Symposium on Interactive 3D Graphics. p. 175–183. I3D '90, Association for Computing Machinery (1990). https://doi.org/10.1145/91385.91442
73. WestEd: The Next Generation Science Standards (2025), https://www.nextgenscience.org/
74. White, R.W., Roth, R.A.: Exploratory Search: Beyond the Query-Response Paradigm. No. 3 in Synthesis Lectures on Information Concepts, Retrieval, and Services, Springer Cham (2009). https://doi.org/10.1007/978-3-031-02260-9
75. Wiliam, D.: What is Assessment for Learning? Studies in Educational Evaluation **37**(1), 3–14 (2011). https://doi.org/10.1016/j.stueduc.2011.03.001
76. Wilton, L., Ip, S., Sharma, M., Fan, F.: Where Is the AI? AI Literacy for Educators. In: Rodrigo, M.M., Matsuda, N., Cristea, A.I., Dimitrova, V. (eds.) Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium. pp. 180–188. Springer International Publishing, Cham (2022)
77. Zamfirescu-Pereira, J., Wong, R.Y., Hartmann, B., Yang, Q.: Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23, Association for Computing Machinery (2023). https://doi.org/10.1145/3544548.3581388

78. Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., Hashimoto, T.B.: Benchmarking Large Language Models for News Summarization. Transactions of the Association for Computational Linguistics **12**, 39–57 (01 2024). https://doi.org/10.1162/tacl_a_00632
79. Zhou, T., Huang, J., Chan, G.Y.Y.: Epigraphics: Message-Driven Infographics Authoring. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. CHI '24, Association for Computing Machinery (2024). https://doi.org/10.1145/3613904.3642172
80. Zylich, B., Viola, A., Toggerson, B., Al-Hariri, L., Lan, A.: Exploring Automated Question Answering Methods for Teaching Assistance. In: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds.) Artificial Intelligence in Education. pp. 610–622. Springer International Publishing, Cham (2020)